

perception network.  
DC P86 T01 U22 W04  
IN TRYBA, V  
PA (SIBE-N) SIBET GMBH SIKAN FORSCHUNGS & ENTWICKLUN; (SICA-N) SIKAN F & E  
GMBH SIBET  
CYC 1  
PI DE 19705471 A1 19970724 (199735)\* 9p G10L005-06 <--  
DE 19705471 C2 19980409 (199818) 9p G10L005-06 <--  
ADT DE 19705471 A1 DE 1997-19705471 19970213; DE 19705471 C2 DE 1997-19705471  
19970213  
PRAI DE 1997-19705471 19970213  
IC ICM G10L005-06  
AB DE 19705471 A UPAB: 19970828

The method involves a neural network with an output magnitude characteristic which is time dependent. The characteristics are extracted from a predefined relation, and the time signal is obtained directly, without expensive frequency transformations, being a result of a learning process in the neural network (1).

For each word to be learnt, a perception (2) is used, adapted to this word. Each perception is a single-layer neural network, and about 50 characteristics are extracted from the speech signal, and adapted to the neural network. The calculation depends on certain functions by the central microprocessor, which are functions from various probability studies.

USE/ADVANTAGE - For speech recognition systems. Requires modest memory capacity, low power consumption and small expenditure on computers. Dwg.1/5

FS EPI GMPI  
FA AB; GI  
MC EPI: T01-C08A; U22-G01A5; U22-G01D; W04-V01

L2 ANSWER 3 OF 10 WPINDEX COPYRIGHT 2000 DERWENT INFORMATION LTD  
AN 1997-109820 [11] WPINDEX  
DNN N1997-090806  
TI Speech signal noise reduction arrangement - corrects speech signals from microphones according to frequency band, using digital recursive low-pass filter and signal-to-noise ratio detector.

DC P86 S05 U22 W01 W04  
IN MARTIN, R  
PA (SIEI) SIEMENS AG  
CYC 3  
PI DE 19524847 C1 19970213 (199711)\* 5p G10L007-04 <--  
JP 09034496 A 19970207 (199716) 4p G10L009-00  
US 5699480 A 19971216 (199805) 5p G10L003-02  
ADT DE 19524847 C1 DE 1995-19524847 19950707; JP 09034496 A JP 1996-173981  
19960703; US 5699480 A US 1996-673151 19960701  
PRAI DE 1995-19524847 19950707  
IC ICM G10L003-02; G10L007-04; G10L009-00  
ICS G10L007-02; H03H017-00; H03H021-00  
AB DE 19524847 C UPAB: 19970313

The arrangement divides the signal processing required for noise correction into three frequency bands. Two microphones are included for communication, e.g. between a patient and a doctor while the patient is being examined in a tomograph. The low frequency band of the microphone signals are high-pass filtered.

The middle frequency band is weighted with a scalar factor so that this frequency band is suppressed during breaks in speech. In the upper frequency band, an adaptive filter is used. The signals are feedback via an inverse filter. A digital recursive low pass filter is included, as is a signal-to-noise ratio detector.

USE/ADVANTAGE - Speech transmission during computer tomography or magnetic resonance measurement. Reduces effect of noise on speech transmission quality.

Dwg. 2/3

FS EPI GMPI  
FA AB; GI  
MC EPI: S05-D02X; U22-G01A1; U22-G01B2; U22-G01B3; U22-G01D; W01-C04A;  
W04-V05E

L2 ANSWER 4 OF 10 WPINDEX COPYRIGHT 2000 DERWENT INFORMATION LTD  
AN 1993-068306 [09] WPINDEX  
DNN N1993-052419  
TI Digital acoustic signal processor esp. for hearing aid - provides choice  
of reprodn. of speech in real time or at speed adapted to hearing defect.  
DC P86 W04  
IN HOTTA, M; IKEDA, H; NEJIME, Y  
PA (HITA) HITACHI LTD  
CYC 3  
PI DE 4227826 A1 19930225 (199309)\* 16p G10L005-02 <--  
JP 05056499 A 19930305 (199315) H04R025-00  
JP 05252594 A 19930928 (199343) H04R025-00  
US 5794201 A 19980811 (199839) G10L003-00  
DE 4227826 C2 19990722 (199933) G10L003-02 <--  
JP 3008640 B2 20000214 (200013) 6p H04R025-00  
ADT DE 4227826 A1 DE 1992-4227826 19920821; JP 05056499 A JP 1991-211872  
19910823; JP 05252594 A JP 1992-45257 19920303; US 5794201 A Div ex US  
1992-931375 19920818, US 1995-462268 19950605; DE 4227826 C2 DE  
1992-4227826 19920821; JP 3008640 B2 JP 1992-45257 19920303  
FDT JP 3008640 B2 Previous Publ. JP 05252594  
PRAI JP 1991-211872 19910823; JP 1992-45257 19920303  
IC ICM G10L003-00; G10L003-02; G10L005-02; H04R025-00  
ICS G10L009-00; H04R003-00; H04R025-02  
AB DE 4227826 A UPAB: 19930924  
The signal picked up by a microphone (2) is amplified and digitised (11)  
for storage in semiconductor memory (14) and processing (12) which  
involves improvement of acoustic properties (121) and low-speed sound  
reproduction (122).  
A controller (4) enables the user to select between real-time and  
time-expanded processing of the stored digital signals before they are  
reconverted to analogue form and amplified (13) to the level required to  
drive an earphone (3).  
ADVANTAGE - Hearing characteristic of elderly person with degraded  
time resolution can be compensated by reproduction of digitised stored  
speech at slower rate. (Dwg. 2/11  
2/11

FS EPI GMPI  
FA AB; GI  
MC EPI: W04-G01B7; W04-V05; W04-V09; W04-Y03

L2 ANSWER 5 OF 10 WPINDEX COPYRIGHT 2000 DERWENT INFORMATION LTD  
AN 1992-277685 [34] WPINDEX  
TI Data reduced speech communication based on non-harmonic constituents -  
involves analogue-digital converter receiving band limited input signal  
with digital signal divided into twenty one band passes at specific time.  
DC P86 U21 W01 W04  
IN KOENIG, F  
PA (KOEN-I) KOENIG F  
CYC 1  
PI DE 4203436 A 19920813 (199234)\* 14p G10L007-04 <--  
ADT DE 4203436 A DE 1992-4203436 19920206  
PRAI DE 1991-4103568 19910206  
IC ICM G10L007-04  
ICS H03M003-00; H03M007-30; H04M001-64  
AB DE 4203436 A UPAB: 19931006  
The signal processing stages appertain to a digital input-side,  
data-reduced speech signal synthesis, i.e., the prodn. of a

*incomplete*

19 BUNDESREPUBLIK  
DEUTSCHLAND



DEUTSCHES  
PATENTAMT

12 **Offenlegungsschrift**  
10 **DE 197 05 471 A 1**

51 Int. Cl.<sup>8</sup>:  
**G 10 L 5/06**

21 Aktenzeichen: 197 05 471.4  
22 Anmeldetag: 13. 2. 97  
43 Offenlegungstag: 24. 7. 97

DE 197 05 471 A 1

Mit Einverständnis des Anmelders offengelegte Anmeldung gemäß § 31 Abs. 2 Ziffer 1 PatG

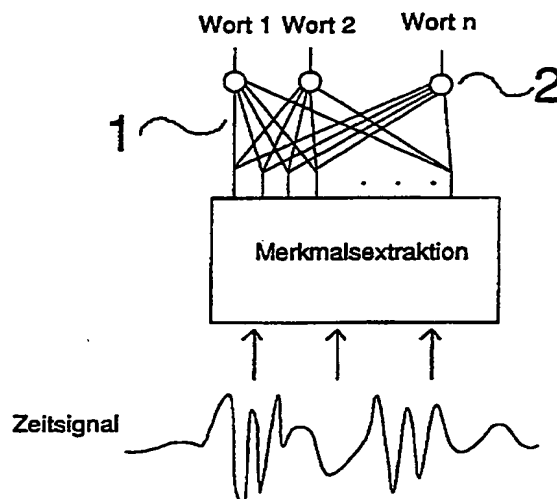
71 Anmelder:  
SIBET GmbH (SICAN Forschungs- und  
Entwicklungsbetriebsgesellschaft), 30419 Hannover,  
DE

72 Erfinder:  
Tryba, Viktor, Dr., 33330 Gütersloh, DE

Prüfungsantrag gem. § 44 PatG ist gestellt

54 Verfahren und Schaltungsanordnung zur Spracherkennung und zur Sprachsteuerung von Vorrichtungen

57 Verfahren und Schaltungsanordnung zur Spracherkennung und zur Sprachsteuerung von Vorrichtungen.  
Zur Spracherkennung werden neuronale Netze verwendet, die sehr rechenaufwendig sind. Die Klassifizierungs-Merkmale müssen relativ aufwendig ermittelt werden, um eine zuverlässige Funktion des Verfahrens zu gewährleisten. Es hat sich gezeigt, daß die Verwendung eines Transversalfilters in Verbindung mit einem Perzeptron-Netz sehr gut zur wenig rechenintensiven Spracherkennung geeignet ist. Durch die Ermittlung einer Einhüllenden und die Aufteilung des Signals in vier Teilbereiche wird der Rechenbedarf noch weiter verringert und die Zuverlässigkeit des Systems gesteigert.  
Die Erfindung kann insbesondere zur preiswerten Sprachsteuerung von Haushaltsgeräten verwendet werden.



DE 197 05 471 A 1

Die Erfindung betrifft ein Verfahren und eine Schaltungsanordnung zur Spracherkennung und zur Sprachsteuerung von Vorrichtungen. Die Erkennung der Kommandoworte erfolgt prinzipiell dadurch, daß Kommandowort-Signale digital aufgezeichnet und Merkmale der digital aufgezeichneten Signale berechnet werden, die in ein neuronales Netzwerk zur Ermittlung des zugehörigen Kommandowortes gespeist werden.

Spracherkennungsverfahren, die auf neuronalen Netzwerken basieren, sind vor allem auf der Basis von verborgenen Markov-Modellen (Hidden Markov-Modell) oder dem Dynamic Time Warping-Verfahren (DTW) bekannt. Diesbezüglich sei auf die DE-PS 33 37 353 C2, DE-OS 42 41 688 A1, DE-OS 195 08 711 A1 und DE 44 17 557 A1 verwiesen. Alle diese Verfahren sind sehr aufwendig und erfordern leistungsfähige digitale Signalprozessoren zur Durchführung einer Frequenzanalyse. Außerdem müssen die Daten für alle Sprachproben aufgezeichnet und permanent verfügbar gehalten werden, damit bei dem späteren Erkennungsvorgang das ähnlichste Wort im Vergleich zu den Sprachproben herausgesucht werden kann. Dementsprechend sind die Anforderungen an die Prozessorleistung und den Speicherbedarf relativ hoch und der Herstellungspreis vor allem zur Anwendung bei Sprachsteuerungen zu hoch.

Der Rechenaufwand ist auch bei der in der DE-OS 41 11 995 A1 beschriebenen Schaltungsanordnung zur Spracherkennung relativ groß, da dort eine Spektralanalyse durchgeführt werden muß.

In der DE-OS 39 31 638 A1 wird ein Verfahren zur sprecheradaptiven Spracherkennung beschrieben, das keine Frequenztransformation mehr erfordert. Aus dem Sprachsignal werden Merkmalsvektoren extrahiert, die in silbenorientierte Wortuntereinheiten segmentiert und klassifiziert werden. Pro Wortuntereinheit wird ein Vergleich mit Referenzmustern durchgeführt. Hierzu müssen die Referenzmuster für einen gesamten Wortschatz abgespeichert und für den Vergleich verfügbar gehalten werden.

In der DE-PS 39 35 308 C1 wird zur Spracherkennung die Durchführung einer Differenzierung und einer Deltamodulation des abgetasteten Sprach-Zeitsignals vorgeschlagen. Anschließend wird die Anzahl der "Einsen" festgestellt, die in Bytes vorhanden sind, die aus einer Anzahl aufeinanderfolgender Abtastwerte gebildet werden. Die Anzahl von "Einsen" pro Byte repräsentiert eine Hauptcodezahl, die über die Zeit aufgetragen ein Balkenmuster bildet, das mit Referenzmustern verglichen wird. Auch hier ist eine Abspeicherung einer Vielzahl von Referenzmustern erforderlich.

In der DE-OS 41 03 913 A1 ist eine Einrichtung zur Gerätesteuerung beschrieben, bei dem ein Zeitmuster in bezug auf die Ton- und Pausendauer ermittelt wird, das mit Referenzmustern verglichen wird. Die Unterscheidungsfähigkeit von Worten allein durch das Merkmal der Ton- und Pausendauer ist in der praktischen Anwendung nicht ausreichend. Außerdem müssen pro Kommandowort Referenzmuster abgespeichert und verfügbar gehalten werden.

In der DE-OS 195 08 137 A1 werden Worte schrittweise klassifiziert, indem eine Teilwortlänge, eine Anzahl von Segmenten und eine Lauttypfolge ermittelt wird. Für den nachfolgenden Klassifikationsschritt werden dann nur noch die Referenzworte betrachtet, bei denen die Merkmale innerhalb vorgegebener Toleran-

zen liegen.

Zur Detektion des Start- und Endpunktes eines Wortes wird in der DE-OS 44 22 545 A1 vorgeschlagen das Sprechsignal blockweise zu unterteilen und Merkmalsvektoren zu bilden, indem pro Block die Signalenergie sowie die quadratische Differenz eines LPC (Linear-Predictive-Coding)-Cepstrum-Koeffizienten in bezug auf einen mittleren LPC-Cepstrum-Koeffizienten bestimmt wird. Der Merkmalsvektor werden mit einem mittleren Merkmalsvektor werden mit einem Schwellwert zur Detektion des Start-/Endpunktes verglichen.

In der DE-OS 43 28 752 A1 wird ein Spracherkennungssystem vorgestellt, das ein mehrschichtiges neuronales Netzwerk erfordert. Dadurch ist der Rechenaufwand und die Anzahl von Netzwerkknöten unerwünscht hoch.

### Aufgabe

Ausgehend von diesem Stand der Technik war es Aufgabe der Erfindung, ein Verfahren und eine Schaltungsanordnung zur Spracherkennung und zur Sprachsteuerung von Vorrichtungen mit vermindertem Rechenaufwand zu schaffen, wobei nur eine geringe Leistung eines zentralen Mikroprozessors (CPU) und eine geringe Speicherkapazität erforderlich ist. Das Verfahren und die Vorrichtung sollte dennoch sehr zuverlässig und leistungsfähig sein.

### Erfindung

Die Aufgabe wird durch das Verfahren nach Anspruch 1 und die Schaltungsanordnung nach Anspruch 15 gelöst.

Vorteilhafte Ausgestaltungen sind in den Unteransprüchen beschrieben.

Es hat sich gezeigt, daß die Verwendung eines Transversalfilters in Verbindung mit einem Perzeptron-Netz sehr gut zur wenig rechenintensiven Spracherkennung geeignet ist. Durch die Ermittlung einer Einhüllenden und die Aufteilung des Signals in vier Teilbereiche wird der Rechenbedarf noch weiter verringert und die Zuverlässigkeit des Systems gesteigert.

Werden bei dem Anlernvorgang Sprechproben von mehreren Personen aufgenommen, ist das Verfahren in der Lage, eine gewisse Sprecherunabhängigkeit bei der Spracherkennung zu erreichen.

### Zeichnungen

Die Erfindung wird nachfolgend anhand der Zeichnungen näher erläutert. Es zeigen:

Fig. 1 Verfahrensprinzip zur Spracherkennung unter Verwendung eines künstlichen neuronalen Netzes;

Fig. 2 Prinzip der Ermittlung der Einhüllenden aus dem Zeitsignal;

Fig. 3 Berechnung der Merkmale jeweils für die vier Wortbereiche;

Fig. 4 Stark vereinfachtes digitales Transversalfilter;

Fig. 5 Schaltungsanordnung zur Spracherkennung.

### Ausführungsbeispiele

Das Verfahrensprinzip zur Spracherkennung unter Verwendung eines neuronalen Netzes ist in der Fig. 1 dargestellt. Das Sprachsignal ist über die Zeitachse aufgetragen.

Das Prinzip des Verfahrens besteht darin, daß Merk-

male zur Erkennung des Wortes ohne aufwendige Frequenztransformation direkt aus dem Zeitsignal extrahiert werden. Die extrahierten Merkmale werden an ein künstliches neuronales Netz (1) angelern. Für jedes anzulernende Wort wird ein Perzeptron (2) verwendet, das auf dieses Wort angelern wird. Ein Perzeptron ist ein einschichtiges neuronales Netz, daß z. B. in "The perceptron: a probabilistic model for information storage and organization in the brain" von F. Rosenblatt in "Neurocomputing: foundations of research", Massachusetts Institute of Technology, 1988, S. 92 ff., ausführlich dargestellt ist. Insgesamt werden ca. 50 Merkmale aus dem Sprachsignal extrahiert und dem neuronalen Netz (1) angelern.

Die Berechnung der Merkmale erfolgt durch Funktionen, die auf dem zentralen Mikroprozessor ablaufen. Dieser ist vorteilhafterweise ein Mikrocontroller, der Schaltungen zur Ansteuerung von Steuerelementen und zur Analog-Digital-Wandlung auf dem Chip aufweist.

Die ermittelten Merkmalsvektoren werden vor dem Anlernen an das neuronale Netz (1) normiert, und zwar in der Weise, daß für jedes Merkmal getrennt das Maximum des Betrages ermittelt wird. Danach wird die betreffende Komponente durch diesen Wert dividiert, womit erreicht wird, daß alle Merkmale in den Zahlenbereich  $-1 \dots +1$  abgebildet werden. Dies führt zu einer Erhöhung der Erkennungsrate, da alle Merkmale gleich stark gewichtet sind und nicht einzelne Merkmale mit einem kleinen Absolutwert gegenüber Merkmalen mit großen Absolutwerten vernachlässigt werden.

Für die neu zu klassifizierenden Merkmale muß entsprechend die Multiplikation jeder Komponente erfolgen.

#### Beschreibung der Merkmale

Im folgenden werden die Verfahrensschritte zur Spracherkennung und zur Ermittlung der Klassifikations-Merkmale für das Sprachsignal beschrieben. Es sind vorgesehen:

1. Ermittlung statistischer Größen;
2. Ermittlung der Einhüllenden zur Extrahierung von Merkmalen;
3. Bestimmung des Durchschnittswerts der Einhüllenden;
4. Detektion der Wortanwesenheit (kein Merkmal);
5. Detektion von Wortanfang und Wortende (kein Merkmal);
6. Bestimmung von Vorläufern und Nachläufern;
7. Bestimmung der Anzahl der Silben;
8. Unterteilung des Wortes in vier Teilbereiche;
9. Anwendung stark vereinfachter digitaler Filter;
10. Bestimmung der Signalenergie nach einer Tiefpaß und Hochpaß-Filterung;
11. Bestimmung der Anzahl der Nulldurchgänge.

#### 1. Ermittlung statistischer Größen

Zunächst werden einige einfache statistische Größen aus dem Zeitsignal berechnet, wie der Mittelwert des Signales, die Varianz, die Gesamtsumme des Signales und die Wortdauer.

#### 2. Ermittlung der Einhüllenden zur Extrahierung von Merkmalen

Zur Einsparung von Rechenzeit wird danach die Ein-

hüllende des Zeitsignals berechnet. Zu diesem Zweck wird, wie in der Fig. 2 skizziert ist, jeweils in einem Teilintervall das jeweilige Maximum ermittelt und gespeichert. Die Gesamtmenge der auszuwertenden Daten reduziert sich dabei von 20 000 Abtastwerten auf ca. 150 Abtastwerte. Diese sind ausreichend, um die Einhüllende hinreichend genau zu beschreiben. Aus der Einhüllenden wird ein Teil der Klassifikations-Merkmale gewonnen.

Aus der Form der Einhüllenden lassen sich dann weitere Merkmale mit vergleichsweise geringem CPU-Aufwand berechnen. Die Verwendung der Einhüllenden für diese Merkmalsbestimmungen macht das Ergebnis zudem robuster gegen einzelne Störsignale und Variationen der Aussprache.

#### 3. Bestimmung des Durchschnittswerts der Einhüllenden

Es wird der Durchschnittswert der Einhüllenden bestimmt. Dieser unterscheidet sich vom Durchschnittswert des Zeitsignales, da entsprechend der Fig. 2 eine Art Gleichrichtung und Glättung stattgefunden hat.

#### 4. Detektion der Wortanwesenheit

Zunächst wird mit der Einhüllenden in robuster Weise detektiert, ob überhaupt ein Wort gesprochen wurde. Zu diesem Zweck wird geprüft, ob der Durchschnittswert der Einhüllenden deutlich und für eine längere Zeit (mind. 0,2 Sekunden) überschritten wurde. Daraus wird zunächst ein Zeitpunkt bestimmt, der als Wortmitte bezeichnet wird.

#### 5. Detektion von Wortanfang und Wortende

Ausgehend von dieser Wortmitte wird sodann der Wortanfang und das Wortende gesucht. Dazu muß ein Schwellwert der Einhüllenden unterschritten werden, und danach für längere Zeit unterschritten bleiben (Stille vor und nach dem Wort).

#### 6. Bestimmung von Vorläufern und Nachläufern

Manche Worte zeichnen sich dadurch aus, daß es Vor- oder Nachläufer gibt. Um dies festzustellen, wird aus der Einhüllenden eine Ableitung bestimmt. Die Beträge der Ableitungen werden im Anfangsbereich und Endbereich des Wortes aufsummiert. Je größer die erhaltenen Werte sind, desto eher kann angenommen werden, daß Vor- bzw. Nachläufer vorhanden sind. Mit diesem Merkmal wird zugleich auch ein Maß für ihre Intensität ermittelt.

#### 7. Bestimmung der Anzahl der Silben

Die Anzahl der Silben eines Wortes kann nicht mit einfachen Algorithmen ermittelt werden, da beispielsweise das Wort "zurück" von manchen Sprechern mit, von anderen ohne Pause gesprochen wird bzw. es auch Übergänge gibt. Um ein Maß dafür zu erhalten, ob es eine Pause in der Wortmitte gibt, werden aus der Einhüllenden Ableitungen bestimmt und die Beträge der Ableitungen im Bereich der Wortmitte aufsummiert.

#### 8. Unterteilung des Wortes in vier Teilbereiche

Mit Hilfe der Einhüllenden kann das Wort in die vier

gleichgroßen Teilbereiche 1. Viertel, 2. Viertel, 3. Viertel, 4. Viertel unterteilt werden. Die Einteilung ist aus der Fig. 3 ersichtlich. Die im folgenden beschriebenen extrahierten Merkmale werden dann jeweils für diese Teilbereiche, auch Abschnitte genannt, berechnet.

Dieses Vorgehen ist sinnvoll, da sich die Eigenschaften des Zeitsignales im Verlaufe der Aussprache eines Wortes ändern. Es hat sich gezeigt, daß eine feinere Unterteilung des Wortes in wesentlich mehr Abschnitte nicht sinnvoll ist, da sich damit die pro Zeiteinheit zu verarbeitende Datenmenge erhöht, sich aber die Robustheit des Erkennungsalgorithmus hingegen verringert, da Einzelheiten des Zeitsignales und zufällige Schwankungen ein zu großes Gewicht erhalten.

#### 9. Verwendung von stark vereinfachten digitalen Filtern

In Lehrbüchern, z. B. in "Einführung in die digitale Signalverarbeitung", H. Götz, Teubner Studienskripten, Stuttgart, 1990, S. 110, wurde gezeigt, daß die FFT (Fast Fourier Transform) im Prinzip als eine Filterbank aus vielen einzelnen Bandpässen verstanden werden kann. Dabei ist der Aufwand für die Realisierung dieser Bandpässe relativ groß.

Der Aufwand kann jedoch deutlich reduziert werden. Im Verlaufe der Entwicklung des Gegenstandes der Erfindung hat sich gezeigt, daß extrem vereinfachte digitale Filter in Kombination mit einem Perzeptron-Netzwerk zu guten Ergebnissen bei der Spracherkennung führen. Zu diesem Zweck wird das folgende, stark vereinfachte digitale Transversalfilter verwendet, das in der Fig. 4 dargestellt ist.

Es wird jeweils die Differenz aus dem aktuellen Abtastwert  $z_t$  des Sprachsignales mit einem um die Zeit  $\tau$  älteren Wert  $z_{t-\tau}$  gebildet:  $d_t = z_t - z_{t-\tau}$ . Die Absolutwerte der Differenz eines Teilbereichs (Wortviertels) werden jeweils für das 1., 2., 3. und 4. Wortviertel aufsummiert und als Merkmal dem Perzeptron zugeführt. Es werden die Merkmale  $S_1, S_2, S_3, S_4$  ermittelt:

$$S_1 = \sum_{t_1}^{t_2} |z_t - z_{t-\tau}|;$$

$$S_2 = \sum_{t_2}^{t_3} |z_t - z_{t-\tau}|;$$

$$S_3 = \sum_{t_3}^{t_4} |z_t - z_{t-\tau}|;$$

$$S_4 = \sum_{t_4}^{t_5} |z_t - z_{t-\tau}|.$$

Durch die Zeitverzögerung und die Differenzbildung ergibt sich ein frequenzselektives Verhalten des Merkmals, das von  $\tau$  abhängt.

Werden unterschiedliche Verzögerungszeiten  $\tau$  gewählt, so ist das Merkmal jeweils für verschiedene Frequenzbereiche sensitiv, so daß eine Frequenzanalyse des Wortes erfolgt, die mit vergleichsweise schlechten

digitalen Filtern auskommt, die aber in Verbindung mit dem Anlernvorgang des neuronalen Netzes dennoch zu guten Erkennungsergebnissen führt.

Insgesamt werden ca. 20 derartige Merkmale aus dem Zeitsignal bestimmt und als Merkmal in das neuronale Netz eingespeist. Die guten Erkennungsraten basieren wesentlich auf diesem Verfahren.

#### 10. Bestimmung der Signalenergie nach einer Tiefpaß-, Hochpaß-Filterung

Ferner wird ein einfacher digitaler Hochpaß und ein Tiefpaß verwendet, der aus der Literatur, z. B. aus "Einführung in die digitale Signalverarbeitung", H. Götz, Teubner Studienskripten, Stuttgart, 1990, hinreichend bekannt ist. Der Ausgangswert des Filters wird nach Betragbildung zur Ermittlung einer "Signalenergie" für jedes Wortviertel aufsummiert.

#### 11. Anzahl der Nulldurchgänge

Die Anzahl der Nulldurchgänge wird für jedes Wortviertel bestimmt und als Merkmal verwendet. Dieses Merkmal gibt Hinweise auf die Tonhöhe.

Die Schaltungsanordnung zur Spracherkennung und Sprachsteuerung ist in der Fig. 5 dargestellt. In einer Wandlerschaltung werden Sprachsignale in analoge elektrische Signale mit Hilfe eines Mikrofons und eines Verstärkers umgewandelt. Mit der Wandlerschaltung ist eine Abtastschaltung zur Erzeugung einer Anzahl digitaler Abtastwerte aus dem analogen elektrischen Signal verbunden. Die digitalen Abtastwerte werden in einem Speicher abgelegt. Ein zentraler Mikroprozessor (CPU) ist zur Ausführung von Befehlsfolgen zur Spracherkennung und zur Steuerung der Schaltungsanordnung vorgesehen. Die Schaltung wird mit einer Taktgeneratorschaltung getaktet. Die Befehlsfolgen zur Spracherkennung werden in einem Speicher, z. B. in einem PROM, fest abgelegt. Eine Schalterkombination ist zur Einstellung binärer Zahlen vorgesehen, wobei die Stellung der Schalterkombination dem Mikroprozessor die Anzahl voneinander unterschiedlicher Kommandoworte anzeigt. Ein Anlernzyklus für die Anzahl Kommandoworte kann mit einem Taster gestartet und gestoppt werden. Während des Anlernzyklus werden Kommandoworte aufgezeichnet und jeweils pro Kommandowort mit Hilfe des Mikroprozessors Referenzmerkmale der digitalen Abtastwerte bestimmt. Die Referenzmerkmale werden in einem SRAM-Speicher abgelegt. Eine Segmentanzeige ist zur Anzeige von Zahlen vorgesehen, die jeweils einem Kommandowort oder dem Betriebsmodus der Schaltungsanordnung, insbesondere des Anlernzyklus oder eines Erkennungszyklus zur Steuerung, entsprechen. In einer bevorzugten Ausführungsform ist die mit der Wandlerschaltung verbundene Abtastschaltung zur Erzeugung einer Anzahl digitaler Abtastwerte aus dem analogen elektrischen Signal ein Bestandteil des zentralen Mikroprozessors (CPU).

Das Zeitsignal des gesprochenen Wortes wird mit einem Mikrophon aufgenommen und mit einer Abtastrate von 10 KHz aufgezeichnet. Die Aufzeichnung ist damit geringfügig besser als Telefonqualität. Die Aufnahmedauer beträgt etwa 2 Sekunden. Der Mikrocontroller beginnt mit der Aufzeichnung erst, nachdem ein Signal, das einen Schwellwert überschreitet, registriert wurde. Durch diese Maßnahme wird etwas Speicherplatz beim Aufzeichnen des Signales eingespart, zudem wartet das System auf die Sprachäußerung. Die Aus-

wertung beginnt erst, nachdem das Signal aufgezeichnet wurde.

#### Patentansprüche

1. Verfahren zur Spracherkennung und zur Sprachsteuerung von Vorrichtungen, wobei ein Sprachsignal aufgezeichnet, digitalisiert und Merkmale des Sprachsignals ermittelt werden und jeweils mit Hilfe eines neuronalen Netzwerkes anhand der ermittelten Merkmale des Sprachsignals das zum Sprachsignal zugehörige Wort bestimmt wird, gekennzeichnet durch

- a) Transversalfilterung des digitalen Sprachsignals für eine Anzahl von Frequenzbereichen;
- b) Ermittlung von Merkmalen  $M_i$  pro Frequenzbereich in Abhängigkeit von den Ergebnissen der Transversalfilterung des jeweiligen Frequenzbereichs;
- c) Bestimmung eines Wortes anhand der Merkmale  $M_i$  mit Hilfe eines Perzeptron-Netzes.

2. Verfahren nach Anspruch 1, gekennzeichnet durch Transversalfilterung des digitalen Sprachsignals mit den Schritten von:

- a) Berechnung einer Anzahl von Differenzen  $d_t$  von jeweils einem aktuellen Abtastwert  $z_t$  mit einem um die Verzögerungszeit  $\tau$  zurückliegenden Abtastwert  $z_{t-\tau}$  für eine Reihe von Zeitpunkten  $t$  des digitalisierten Sprachsignals;
- b) Berechnung jeweils der Absolutwerte der Anzahl von Differenzen;
- c) Bildung der Summe  $S_i$  der Absolutwerte der Anzahl von Differenzen  $d_i$ ;

wobei jede Summe  $S_i$  ein Merkmal  $M_i$  für das Perzeptron-Netz ist.

3. Verfahren nach Anspruch 2, gekennzeichnet durch Ausführung der Transversalfilterung für eine Reihe von Verzögerungszeiten  $\tau$ .

4. Verfahren nach einem der vorhergehenden Ansprüche, gekennzeichnet durch Bestimmung der Einhüllenden des Sprachsignals, wobei jeweils in einem Teilintervall das jeweilige Maximum ermittelt und gespeichert wird.

5. Verfahren nach Anspruch 4, gekennzeichnet durch Bestimmung des Durchschnittswertes der Einhüllenden.

6. Verfahren nach einem der vorhergehenden Ansprüche, gekennzeichnet durch Ermittlung des Mittelwertes des Signals, der Varianz, der Gesamtsumme des Signales und der Wortdauer.

7. Verfahren nach einem der vorhergehenden Ansprüche, gekennzeichnet durch Einteilung des Sprachsignals in vier Teilbereiche.

8. Verfahren nach Anspruch 7, gekennzeichnet durch digitale Hochpaß- und Tiefpaßfilterung jeweils der Teilbereiche des Sprachsignals, wobei der Ausgangswert des Filters für jeden Teilbereich aufsummiert wird.

9. Verfahren nach einem der vorhergehenden Ansprüche, wobei eine Prüfung erfolgt, ob der Durchschnittswert der Einhüllenden deutlich und für eine festgelegte Mindestzeit überschritten wurde, um zu erkennen, ob ein Wort gesprochen wurde.

10. Verfahren nach einem der vorhergehenden Ansprüche, gekennzeichnet durch Bestimmen der Wortmitte durch Halbierung der Zeit bestimmt, in der der Durchschnittswert der Einhüllenden deut-

lich und für eine festgelegte Mindestzeit überschritten ist, und Verwenden der Wortmitte als Merkmal für das Perzeptron-Netz.

11. Verfahren nach einem der vorhergehenden Ansprüche, gekennzeichnet durch Bestimmen des Wortanfangs und des Wortendes durch Vergleich der Einhüllenden mit einem Schwellwert, wobei bei einem Wortende der Schwellwert eine festgelegte Zeit unterschritten sein muß.

12. Verfahren nach einem der vorhergehenden Ansprüche, gekennzeichnet durch Erkennen von Vor- oder Nachläufern durch Bestimmung von Ableitungen aus der Einhüllenden und Aufsummierung der Beträge der Ableitungen im Anfangsbereich und Endbereich des Wortes, wobei ein Vor- bzw. Nachläufer vorhanden ist, wenn ein festgelegter Wert überschritten ist, und Verwenden der Existenz und der Intensität der Vor- und Nachläufer als Merkmal für das Perzeptron-Netz.

13. Verfahren nach einem der vorhergehenden Ansprüche, gekennzeichnet durch Bestimmen der Anzahl von Silben eines Wortes durch Berechnung von Ableitungen der Einhüllenden und Aufsummierung der Beträge der Ableitungen im Bereich der Wortmitte und Verwenden der Anzahl von Silben als Merkmal für das Perzeptron-Netz.

14. Verfahren nach einem der vorhergehenden Ansprüche, gekennzeichnet durch Bestimmen einer Anzahl der Nulldurchgänge für jedes Wortviertel und Verwenden der Anzahl der Nulldurchgänge als Merkmal für das Perzeptron-Netz.

15. Schaltungsanordnung zur Sprachsteuerung von Vorrichtungen mit einer Wandlerschaltung zur Umwandlung von Sprachsignalen in analoge elektrische Signale, einer mit der Wandlerschaltung verbundenen Abtastschaltung zur Erzeugung einer Anzahl digitaler Abtastwerte aus dem analogen elektrischen Signal, einem Speicher für die digitalen Abtastwerte, einer Taktgeneratorschaltung, einem zentralen Mikroprozessor (CPU) zur Ausführung von Befehlsfolgen zur Spracherkennung und einem Speicher für die Befehlsfolgen zur Spracherkennung, wobei der zentrale Mikroprozessor mit den Schaltungen und Speichern zur Ansteuerung und Datenübertragung verbunden ist, gekennzeichnet durch

eine Schalterkombination zur Einstellung binärer Zahlen, wobei die Stellung der Schalterkombination dem Mikroprozessor die Anzahl voneinander unterschiedlicher Kommandoworte anzeigt, einem Taster zum Starten und Stoppen eines Anlernzyklus, in dem Kommandoworte aufgezeichnet und jeweils pro Kommandowort mit Hilfe des Mikroprozessors Referenzmerkmale der digitalen Abtastwerte bestimmt werden, einen fest programmierbaren Speicher zur Speicherung der Referenzmerkmale.

16. Schaltungsanordnung nach Anspruch 15, gekennzeichnet durch eine Segmentanzeige zur Anzeige von Zahlen, die jeweils einem Kommandowort oder dem Betriebsmodus der Schaltungsanordnung, insbesondere des Anlernzyklus oder eines Erkennungszyklus zur Steuerung, entsprechen.

17. Schaltungsanordnung nach einem der Ansprüche 15 oder 16, dadurch gekennzeichnet, daß die mit der Wandlerschaltung verbundene Abtastschaltung zur Erzeugung einer Anzahl digitaler Abtastwerte aus dem analogen elektrischen Signal ein

Bestandteil des zentralen Mikroprozessors (CPU)  
ist.

Hierzu 3 Seite(n) Zeichnungen

5

10

15

20

25

30

35

40

45

50

55

60

65

- Leerseite -

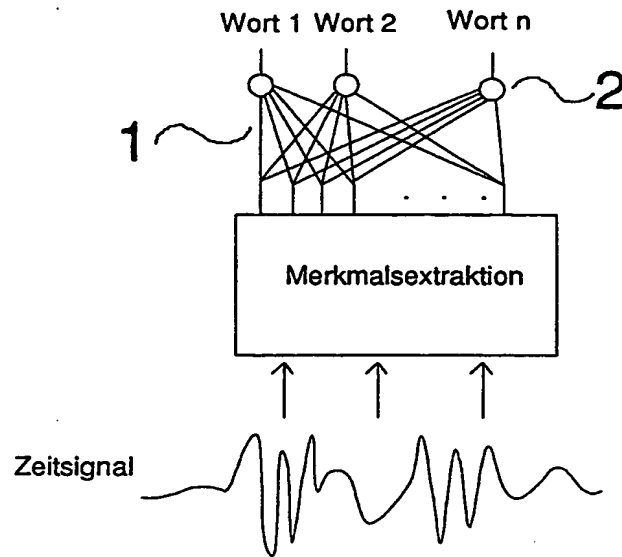


Fig. 1

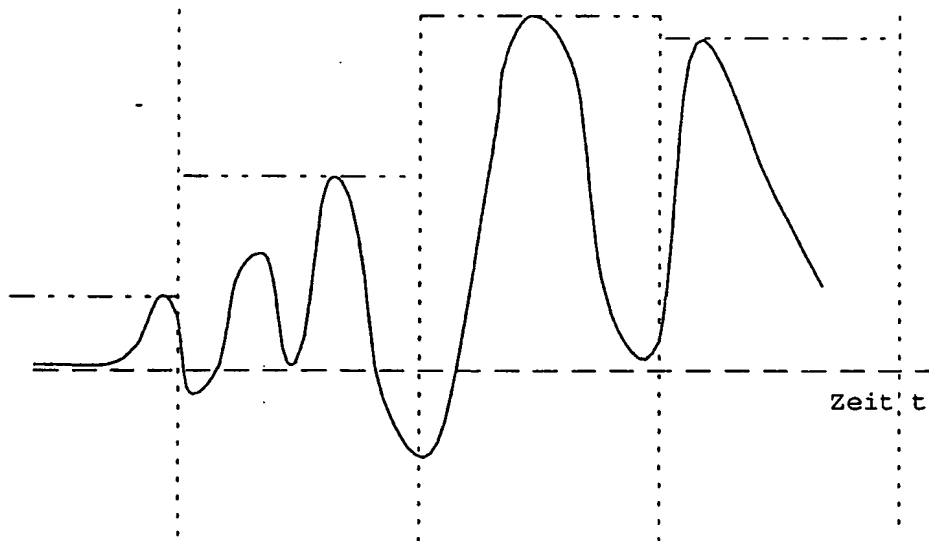


Fig. 2

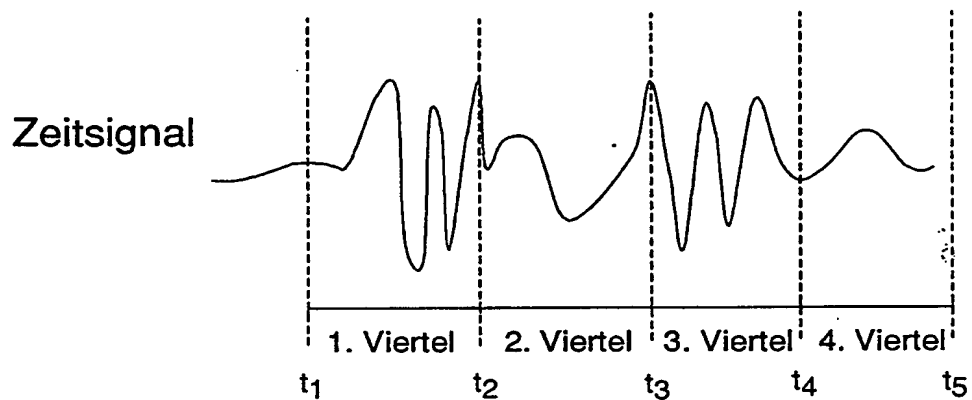


Fig. 3

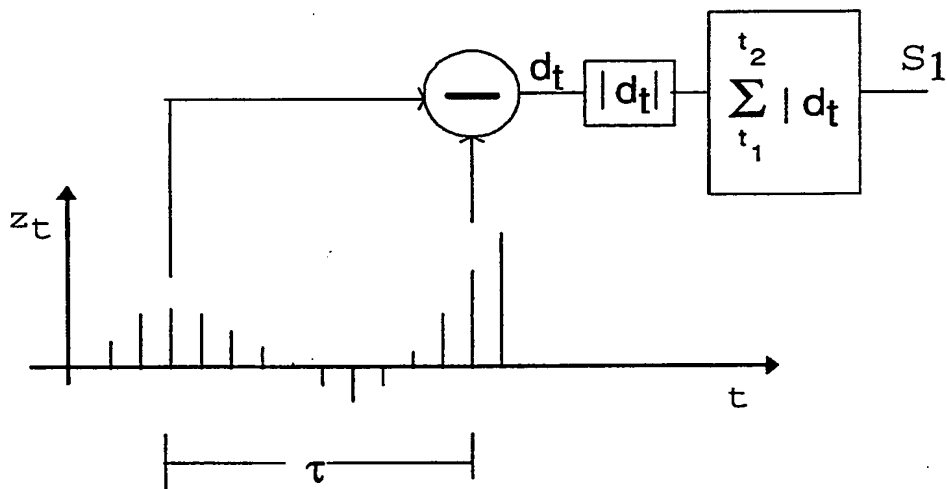


Fig. 4

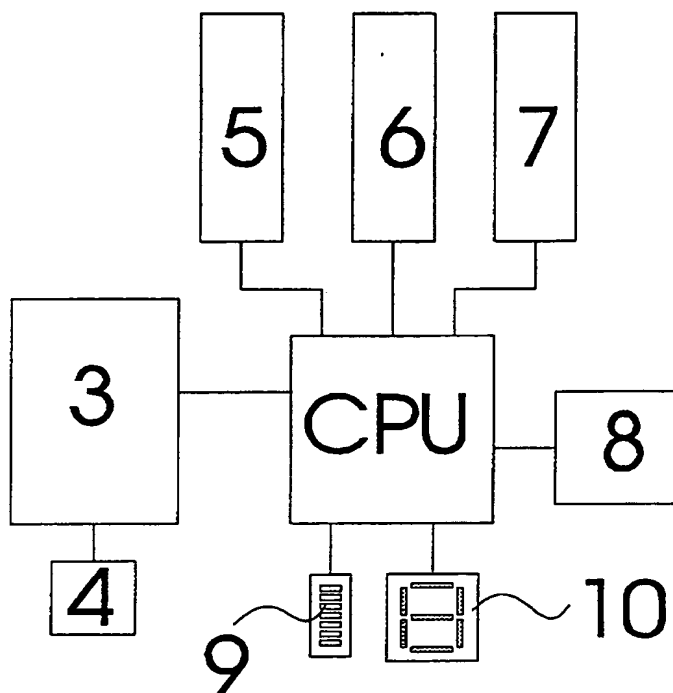


Fig. 5